RELC Journal 25, 2 (1994)
34 – 50

HOW USEFUL IS EAP VOCABULARY FOR ESP? A CORPUS BASED CASE STUDY

CUCU SUTARSYAH
University of Lampung
Indonesia

PAUL NATION and GRAEME KENNEDY
Victoria University of Wellington
New Zealand

This study compares the vocabulary of a single Economics text of almost 300,000 running words with the vocabulary of a corpus of similar length made up of a variety of academic texts. It was found that the general academic corpus used a very much larger vocabulary than the more focused Economics text. A small number of words that were closely related to the topic of the text occurred with very high frequency in the Economics text. The general academic corpus had a very large number of low frequency words. Beyond the words in West's *General Service List* and the University Word List, there was little overlap between the vocabulary of the two corpora. This indicates that as far as vocabulary is concerned, EAP courses that go beyond the high frequency academic vocabulary are of little value for learners with specific purposes.

In the last decade increasing use has been made of machine readable corpora to improve descriptions of the grammar and vocabulary of English. Nowadays learners dictionaries, grammars, and teaching materials are often able to claim to be based on or informed by corpus analysis. Projects are under way to compile and use corpora of English consisting of up to 100 million words or more. The software being developed to analyze such corpora promises to greatly increase our understanding of how likely words and structures are to be used in different subject fields or genres.

Where English is taught for academic purposes (EAP) classes often have to be made up of learners who are interested in different academic fields. Such classes are taught from a range of material arranged thematically or notionally in the expectation that the learners will acquire relevant experience of vocabulary, grammar, and discourse to be used for more specific purposes.

The present corpus study was undertaken to determine the extent to which the vocabulary used in general academic English resembles the vocabulary used in one specialised academic field, namely Economics. Although the two corpora which were compared are small ones, the study can indicate the kinds of insights which corpus analysis and corpus comparison can provide for applied linguistics purposes.

The Two Corpora

The two corpora which were studied were a single university text book, *Macroeconomics* by Michael Parkin (Addison-Wesley, Mass. 1990), and a corpus of a similar length consisting of 160 different texts, each about 2,000 words long, taken from a variety of academic fields including natural sciences, medicine, mathematics, social sciences, political science, law, education, humanities, and technology and engineering. This corpus consists of Section J (Learned and Scientific texts) from both the LOB corpus of written British English (Johansson, Leech and Goodluck, 1978) and the parallel Wellington corpus of written New Zealand English (Bauer, 1993). Table 1 summarises the differences between the two corpora. For convenience, the corpus consisting of Parkin's Economics text is called the Economics text and the collection of 2,000 word academic texts is called the general academic corpus.

Table 1: The Features of the Two Corpora

The Economics text	The general academic corpus
One writer	160 different writers
One topic - macroeconomics	160 different topics
One subject area - economics	At least 15 subject areas
One long continuous text	160 separate 2,000 word texts

In this study the Economics text is taken to represent an English for Specific Purposes (ESP) text and the general academic corpus is taken to represent an English for Academic Purposes (EAP) collection. EAP courses prepare learners for a range of academic disciplines, often by concentrating on the skills and language that are common to these disciplines. ESP courses are more narrowly focused on one academic discipline or one area of expertise or employment. In this paper ESP is regarded as focusing on one academic discipline.

For the study, the publisher's permission was obtained to digitize the Economics text by scanning it into a computer database. The general academic corpus was compiled by extracting the relevant sections from the LOB and Wellington corpora. The analysis was undertaken using the Oxford Concordance Program (OCP) and VocabProfile (a program developed at the English Language Institute, Victoria University of Wellington).

The Goals of the Study

By comparing the vocabulary of the two corpora it is possible to answer the following questions.

- How large a vocabulary is needed to be familiar with most of the vocabulary in an Economics text?
- Does the practice of getting learners to work through a series of unrelated texts as in an EAP course impose too high a vocabulary burden on them?
- What are the values of an EAP course for learners with a specific purpose?
- When should learners move from EAP to ESP?
- How important is the technical vocabulary in a discipline specific text?
- What should an ESP teacher do about technical vocabulary?

This study is largely concerned with the counting and tabulation of vocabulary. This involves turning coherent texts into lists of seemingly unrelated words. Much of what makes up a text is lost in the process, but as we shall see, the information gained about the vocabulary load of the texts can provide insights into aspects of the task that faces teachers and learners when dealing with academic texts.

The assumption behind this study is that learners is ESP and EAP courses need to increase their vocabulary to a level such that the unknown vocabulary in the texts they read does not become a major barrier to comprehension. Successful comprehension involves much more than being able to decode the vocabulary in a text, but a lack of familiarity with more than 5% of the running words in a text can make reading a formidable task (Laufer, 1989). EAP and ESP courses need to ensure that attention is given to vocabulary development and this can be done in a range of ways from direct teaching and learning to incidental learning from extensive reading.

The Vocabulary Size of the Two Corpora

The Economics text and the general academic corpus were compared in terms of length (the number of different tokens), the number of different types or word forms, and the number of different word families. Table 2 contains the results. A word family includes the base word plus inflected and derived forms. For this study the description of word families at Level 4 of Bauer and Nation's (1993) scale was used. This includes plural, third person singular present tense, past tense, past participle, *-ing*, comparative, superlative, possessive; *-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-; -al, -ation, -ess, -ful, -ism, -ist, -ty, -ize, -ment, in-*, all with restricted uses. So in the word family count, for example, *govern, governs, governed, governing, governable, ungovernable, governess, government,* and *governments* were all counted as members of the same word family if they occurred in the corpora. In the rest of this paper, the term *word* means "word families".

Table 2: Text Size and Vocabulary Size of the Two Corpora

Corpus	Tokens	Types	Families
Economics text	295,294	9,469	5,438
General academic corpus	311,768	21,399	12,744

In spite of the corpora being roughly the same length, 295,294 words and 311,768 words, there is a striking difference in their vocabulary size with the general academic corpus containing well over twice as many word families as the Economics text. This means that if the vocabulary of the general academic corpus represents the vocabulary to be learned in an EAP course then the learners would be faced with an enormous vocabulary load.

Let us now look more closely at the occurrence of vocabulary in the two corpora to see more precisely where the differences between them lie.

Low Frequency Words

Table 3 shows the number of word families occurring once, twice, etc, up to 20 times in the two corpora. At all of these low frequency levels of 1 to 20 occurrences the general academic corpus has more word families than the Economics text.

Table 3: The Distribution of Low Frequency Words in the Economics Text and the General Academic Corpus

Frequency	Number of families in the Economics text	Number of families in the general academic corpus
1	1,925	5,364
2	640	1,769
3	397	900
4	232	596
5	170	454
6	151	333
7	108	254
8	108	222
9	80	186
10	69	169
11	79	126
12	71	110
13	51	91
14	54	99
15	41	88
16	40	69
17	36	73
18	36	70
19	37	52
20	36	59

It is clear from Table 3 that the big difference in word families between the two corpora occurs in the low frequency words. This means that learners working their way through the general academic corpus would be continually meeting words that they would not meet again or would meet only a few times in that corpus.

High Frequency Words

The difference between the general academic corpus and the Economics text changes dramatically in the high frequency words. Up to the words occurring 300 times, the general academic corpus almost always has more words. After the 300 word frequency level, the Economics text has more items. Table 4 summarises the data.

Table 4: The Distribution of High Frequency Words in the Economics text and the General Academic Corpus

Frequency range	Economics text	General academic corpus
40 times or less	4,670	11,742
40 to 100 times	348	580
101 to 200 times	194	251
201 to 300 times	71	72
301 to 500 times	63	46
501 to 1,100 times	56	27
1,101 to 24,000 times	36	26

Although in Table 4 the change from the 300 frequency level up does not seem great, it represents a very significant change. Let us see why.

Table 5: The most Frequent 50 Words from the Economics Text and General Academic Corpus

Rank	Economics corpus		General academic corpus	
1	the	22905	the	23890
2	of	12710	of	14591
3	be	10686	be	14021
4	a	9952	and	8455
5	and	8323	in	8077
6	in	7010	to	7867
7	to	6502	a	7857
8	that	4392	that	3400
9	price	3080	has	3217
10	for	2912	this	3143
11	it	2674	it	3071
12	we	2534	for	3060
13	has	2514	as	2574
14	cost	2251	by	2351
15	by	2034	with	2110
16	this	2003	they	2056
17	demand	1944	on	1956
18	on	1882	which	1631
19	as	1831	he	1590
20	they	1820	not	1544
21	curve	1804	or	1542
22	at	1797	at	1535
23	firm	1743	from	1518
24	supply	1590	can	1242
25	quantity	1467	we	1205
26	can	1442	but	1103
27	margin	1427	use	974
28	will	1378	one	922
29	economy	1353	some	894
30	from	1337	there	894
31	produce	1237	do	793
32	if	1202	more	744
33	income	1183	other	704
34	do	1165	all	694

35	with	1135	than	666
36	market	1101	if	662
37	but	1090	make	656
38	or	1058	may	636
39	each	1038	I	615
40	labour	1004	only	608
41	increae	1002	no	595
42	consume	995	when	581
43	than	977	year	577
44	not	974	so	571
45	more	963	such	558
46	other	957	would	556
47	total	946	who	550
48	change	927	two	546
49	rate	915	also	529
50	when	910	any	528

Table 5 lists the 50 most frequent words in both corpora. The 50 most frequent words in the Economics text account for 49.5% of the running words in the corpus. 18 out of these 50 words (36%) are content words. In the general academic corpus, the 50 most frequent words account for 45% of the running words in the corpus. Only 3 of these 50 words *(use, make, year)* are content words (nouns, verbs, adjectives or adverbs).

Many of the high frequency content words in the Economics text also occur in the general academic corpus, but with a lower frequency.

Table 6 lists the words in the 1,000 most frequent words in the Economics text that also occur in the general academic corpus but that occur much more frequently in the Economics text.

Table 6: Words in the First 1,000 of the Economics Text that Occur much more Frequently than in the General Academic Corpus

Word family	Frequency in the Economics text	Frequency in the general academic corpus
Price	3080	90
Cost	2251	91
Demand	1944	102
Curve	1804	83
Firm	1743	41
Supply	1590	86
Quantity	1467	53
Margin	1427	24
Economy	1353	172
Income	1183	96
Produce	1237	167
Market	1104	110
Consume	955	70
Labour	1004	131
Capital	907	50
Total	946	114
Output	861	50
Revenue	763	10
You	866	118
Profit	733	27
Production	772	84
Average	777	90
Goods	705	21
Product	749	106
Trade	621	85
Buy	521	35
Wage	522	75
Monopoly	454	13
Percent	450	41
Million	445	42
Household	360	41
Equilibrium	328	21
Choice	339	39
Elasticity	333	34
Total frequency	**34,594**	**2,322**

The words in the list share several important features. First, with the exception of *you* which reflects a style of writing, all the words are content words clearly related to Economics. Second, *price* is one of the 10 most frequent words in the corpus, and most of them are in the 100 most frequent words in the Economics text. The topic of the text obviously has a major effect on what words occur very frequently. Third, the 34 words in Table 6 occur so frequently that over one in every ten running words used in the Economics text is from this list of 34 words. The general academic corpus does not have a comparable list. Exactly the same vocabulary accounts for less than 1% of the running words in the general academic corpus. The specialised focus of the Economics text thus has a major effect on the occurrence of a small group of content words. These very high frequency words are extremely important for readers of Economics texts, but probably more significant is that the continual use of these words apparently reduces the need to use the more diverse vocabulary that is seen in the academic corpus. Flowerdew (1993) found a similar group of words in a biology corpus. The top nouns in terms of frequency were *cell, water, membrane, food, plant, root, molecule, wall, energy, concentration, organism, cytoplasm, animal, stem, structure, body, part.* Fourth, almost half of the 34 words from Table 7 are in the first 1,000 of the *General Service List of English Words* (West, 1953) and four-fifths are in the GSL or University Word List.

Overlap of the Economics Text and the General Academic Corpus

We have seen the striking difference in vocabulary size between the Economics text and the general academic corpus, and the occurrence of some very high frequency words in the Economics text. These give us some indication of the task facing a learner of English when preparing to learn the vocabulary necessary for access to these corpora. We will now look to see how well the general academic corpus prepares a learner for the specialized Economics text. This is another way of saying, "How well does EAP prepare for ESP?"

The general academic corpus contains at least 7,306 word families (57% of the total 12,744 in the general academic corpus) that do not occur in the Economics text. An EAP course could result in learners working on a lot of vocabulary that is of little immediate use to them in their field of study. This should not be overstated, because general academic courses

have many other goals besides preparing learners for reading in one specialised field. However, most EAP courses operate under severe time restrictions and it clearly seems much more efficient to get learners focused on the language of their particular subject as soon as possible after they are in control of the high frequency general service and academic vocabulary.

2,124 (39%) of the Economics text's 5,438 word families do not occur in the general academic corpus, and almost all of these words (2,026) are not in the general service and general academic vocabulary discussed below. It shows that an EAP course would not prepare learners for much of the vocabulary they would meet in an ESP text in the field of Economics.

General Service and General Academic Vocabulary

So far we have looked only at words according to their frequencies in the two corpora. The design of the vocabulary component of an English course, however, will usually rely on counts based on a much wider sampling of texts. Michael West's (1953) *General Service List of English Words* (GSL) is the classic collection of the high frequency wide range words of English. It consists of a list of the 2,000 most frequent words in general English (words with a wide range of occurrence) with frequency figures for each word and with the percentage of each major meaning of each word indicated. In spite of its age (it was developed much earlier than 1953) and a few errors, it is still a useful list (Hwang and Nation, in press). The vocabulary in this 2000 word list is an essential component of any initial English course. For learners then going on to academic study, the next vocabulary to learn is the 800 word University Word List (UWL) (Xue and Nation, 1984; also in Nation, 1990). This list assumes learners already know the 2,000 word families of the General Service List, and contains words that occur frequently over a range of academic texts, as the following words in the University Word List show, *margin, economy, income, labour, consume, vary, construct, defect, interval.*

Table 7 shows the coverage of the first one thousand and second one thousand words of the GSL and the words of the UWL. For example, 77.72% of the running words in the Economics text are words from the first 1,000 of the GSL.

Table 7: Number of Word Families and Percentage of Coverage of the Economics Text and the General Academic Corpus by GSL and UWL

Word level	Families in the Economics text	Coverage of the Economics text	Families in the general academic corpus	Coverage of the general academic corpus
1st 1,000	1,029	77.72%	1,095	74.11%
2nd 1,000	548	4.78%	796	4.32%
UWL	636	8.74%	811	8.40%
Others	3,225	8.77%	10,042	13.16%
Total	**5,438**	**100%**	**12,744**	**100%**

Table 7 shows that the general academic corpus makes use of a larger number of word families at all word levels including the first 1,000 words. There is a difference of coverage at the 1,000 word level, because of the small group of high frequency content words in the Economics text mentioned above.

A similar effect is also seen at the second 1,000 and UWL levels. The Economics text uses fewer word families at each of these levels, but has slightly greater text coverage. Table 7 shows that subtechnical vocabulary (represented by the UWL) is just as important in a specialised academic text as it is in a diverse range of academic texts. This is important information for a teacher planning the vocabulary component of a course of study for learners with a specialist academic interest. It is also clear that not all of the UWL is important for such learners.

A major difference between the two corpora of course exists in the words not in any of the three lists - 'Others' in Table 6. Fewer words in the Economics text are outside the GSL and UWL. The percentage coverage provided by these words differs significantly, but the difference is not as great as the difference in the actual number of families.

The coverage provided by the GSL and UWL shows that the vocabulary in these lists is of critical importance for learners aiming at academic study. Thus although we saw in the previous section that there was a large amount of vocabulary not common in the Economics text and the general academic corpus, there is still value in an EAP course that covers most of the vocabulary of the UWL. Beyond that however there is little value, from a vocabulary point of view, in EAP for learners whose more specific purposes are already clear to them.

Technical Vocabulary

Technical vocabulary is vocabulary that seems to be particular to and useful for a specific area of knowledge. It is possible to describe it according to frequency and range (Bečka, 1972). From this point of view it can be described as vocabulary whose frequency of occurrence in a particular text or discipline is much higher than its relative frequency over a range of texts or disciplines (Nation, 1990). That means, we find that vocabulary most often in specialised texts than in texts outside that area. To find the useful technical vocabulary in the Economics text, it is necessary to find the following groups of words.

1. Words that occur much more frequently in the Economics text than they do in the general academic corpus

2. Words that occur frequently in the Economics text and which do not occur in the general academic corpus

The words found in this way were checked against a dictionary of Economics (Pearce, 1992). It was estimated that the Economics text contained around 460 technical words. Typically the advice given to teachers of English regarding technical words is "Learning the subject involves learning the [technical] vocabulary. Subject teachers can deal with the [technical] vocabulary but the English teacher can help with learning strategies" (Nation 1990:19). However, the findings of this study show that the more influential of these words (see the 34 items in Table 6) are words from the GSL and UWL. ESP teachers could usefully work with learners on these high frequency technical words, perhaps drawing attention to their generally narrower meaning and pointing out the parts of their meaning that are important for their use in the specialised text. Flowerdew (1993: 236) considers the high frequency words that occur less frequently in other disciplines to be subtechnical vocabulary and deserving attention because "these words are not likely to be glossed by the content teacher".

Using the criteria of frequency and range was only partly successful in isolating the technical words. Other very frequent words in the Economics text and not in the general academic corpus included *Swanky, Jane, haircut, sweater, Pioneerland,* and several others which were used in the Economics text as recurring examples of people, places, products, and services involved in economic activity. There were also technical words in the text which were very infrequent. It may be necessary to define the group of technical words in the same way as high frequency words and general academic vocabulary have been defined in corpus studies, that is by reference to a predetermined list of words, such as a dictionary of Economics.

Conclusion

The following conclusions can be drawn from this study:

1 To know most of the vocabulary in a specialised text a learner of English would need a vocabulary of at least 4,000 to 5,000 words. In fact, learners would have to come to the text with a larger vocabulary than this as it would not be possible to predict without analysis of the text what vocabulary would occur. These must include most of the 2,000 word families of the GSL and the 800 of the UWL.

2 The GSL and UWL are of great value in providing coverage of the vocabulary in both a specialised text and a set of unrelated academic texts. These lists provide around 90% coverage of the words in the text. There is a need for courses that focus on the vocabulary of these two important lists.

3 There is value in the ESP teacher giving some attention to the high frequency technical vocabulary that occurs in some form in other disciplines.

4 A coherent text by a single writer on a single broad topic uses a very much smaller vocabulary than a series of unrelated texts. There are also considerable areas of non-overlap between the vocabulary of such texts. These findings call into question the value of an EAP course for learners who already have specific purposes. Learners with specific purposes should have ESP courses as soon as they are familiar with most of the vocabulary of the GSL and UWL.

5 Most English courses make use of a series of unrelated texts. This can increase the vocabulary load of the course enormously. If teachers or course designers wish to avoid this, it is worth considering making the course consist of a few themes so that the texts within a theme bear more relationship to each other and thus make use of a smaller vocabulary. If the course has a strong fluency focus, it may be useful to narrow the content focus even more by considering the use of one coherent text, at least for that part of the course. This would greatly reduce the vocabulary knowledge needed to cope with the material and allow learners to give their attention to skill development.

6 A corpus based analysis can provide valuable insights which are of use for language pedagogy and course design.

Vocabulary is only one component of a course, but it is a component that learners notice and that can occupy a lot of their learning time. It is a component that deserves more attention from course designers and the aim of this study has been to inform that attention.

References

Bauer, L. 1993. *Manual of Information to Accompany the Wellington Corpus of Written New Zealand English.* Department of Linguistics: Victoria University of Wellington.

Bauer, L. and I. S. P. Nation. 1993. Word Families. *International Journal of Lexicography,* 64, 253-279.

Bečka, J. V. 1972. The lexical composition of specialised texts and its quantitative aspect. *Prague Studies in Mathematical Linguistics*, 4, 47-64.

Flowerdew, J. 1993. Concordancing as a tool in course design. *System*, 21, 2, 231-244.

Hwang, K. and I.S.P. Nation (In press). Where should general service vocabulary stop and special purposes vocabulary begin? *System*, 23, 1.

Johansson, S., G.N. Leech, and H. Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English.* Department of English: University of Oslo.

Laufer, Batia 1989. What percentage of text-lexis is essential for comprehension? In *Special Language: From Humans Thinking to Thinking Machines*. C. Lauren and M. Nordman (eds.), Clevedon: Multilingual Matters.

Nation, I.S.P 1990. *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.

Pearce, D.W. 1992. *Macmillan Dictionary of Modern Economics*. London: Macmillan.

West, M. 1953. *A General Service List of English Words*. London: Longman.

Xue, G. and I.S.P. Nation. 1984. *A University Word List. Language Learning and Communication, 3, 2, 215-229.*